Distillation and Amalgamation

Dr. Ali Arsanjani

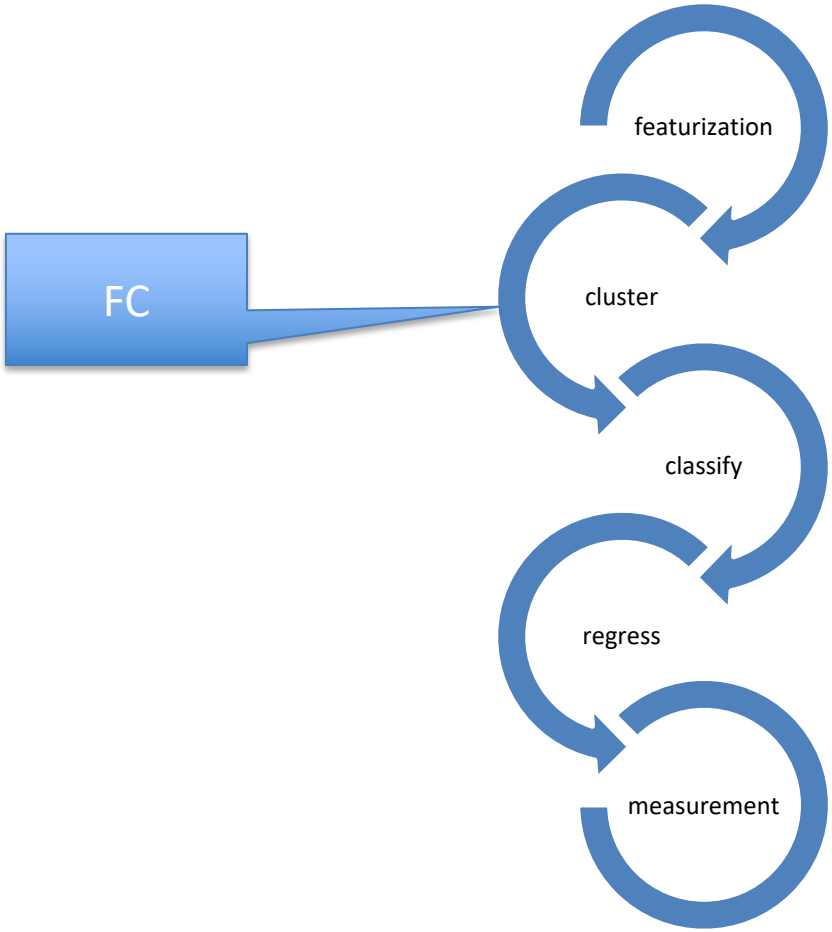
Machine Learning Week 7 (c)
DeepContext 2016-2019

10/8/20

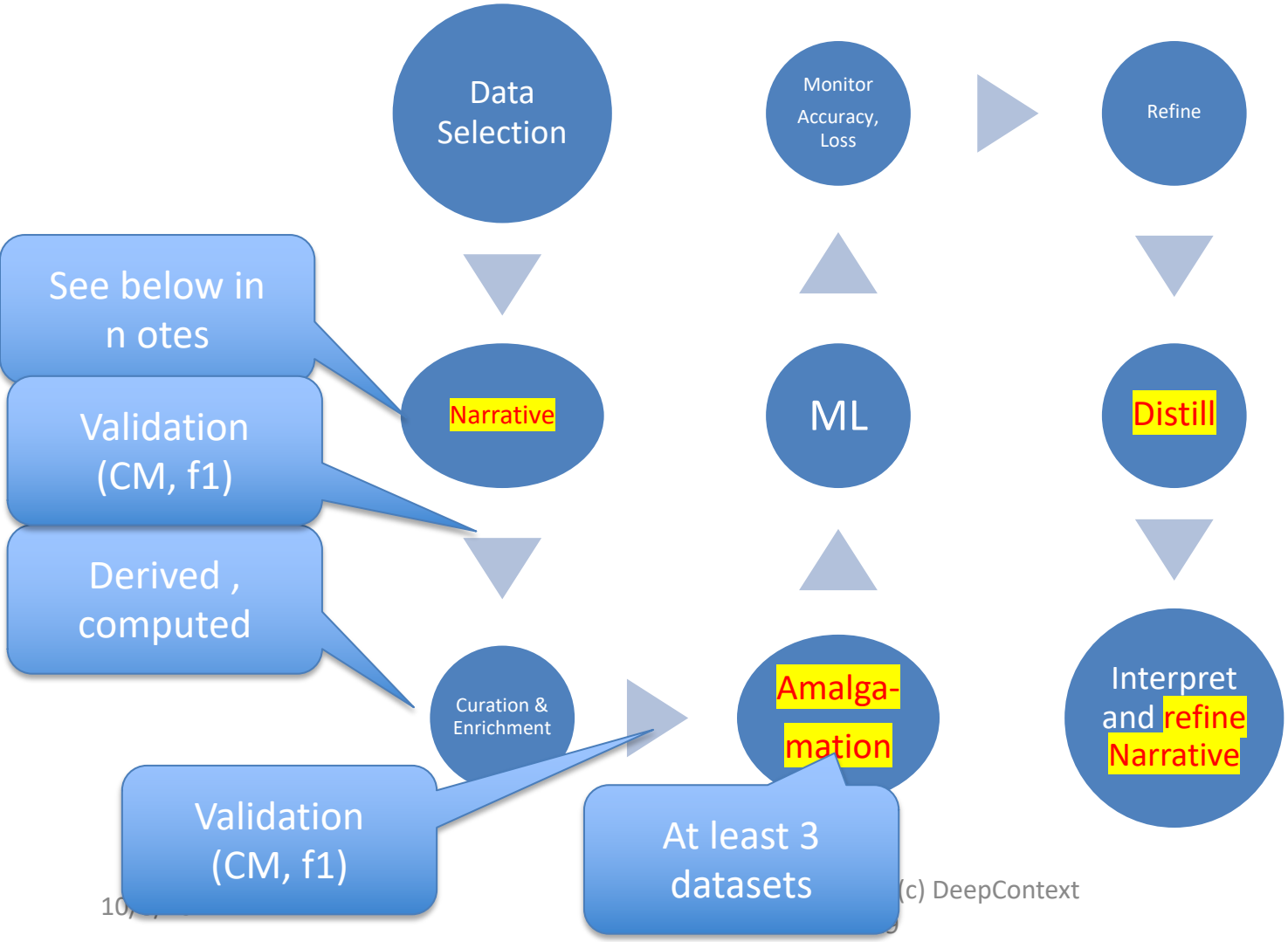
ML

- Clustering, Classification, regression
- Recommendations
- Forecasting (time-series)
  - ARIMA, unique algos

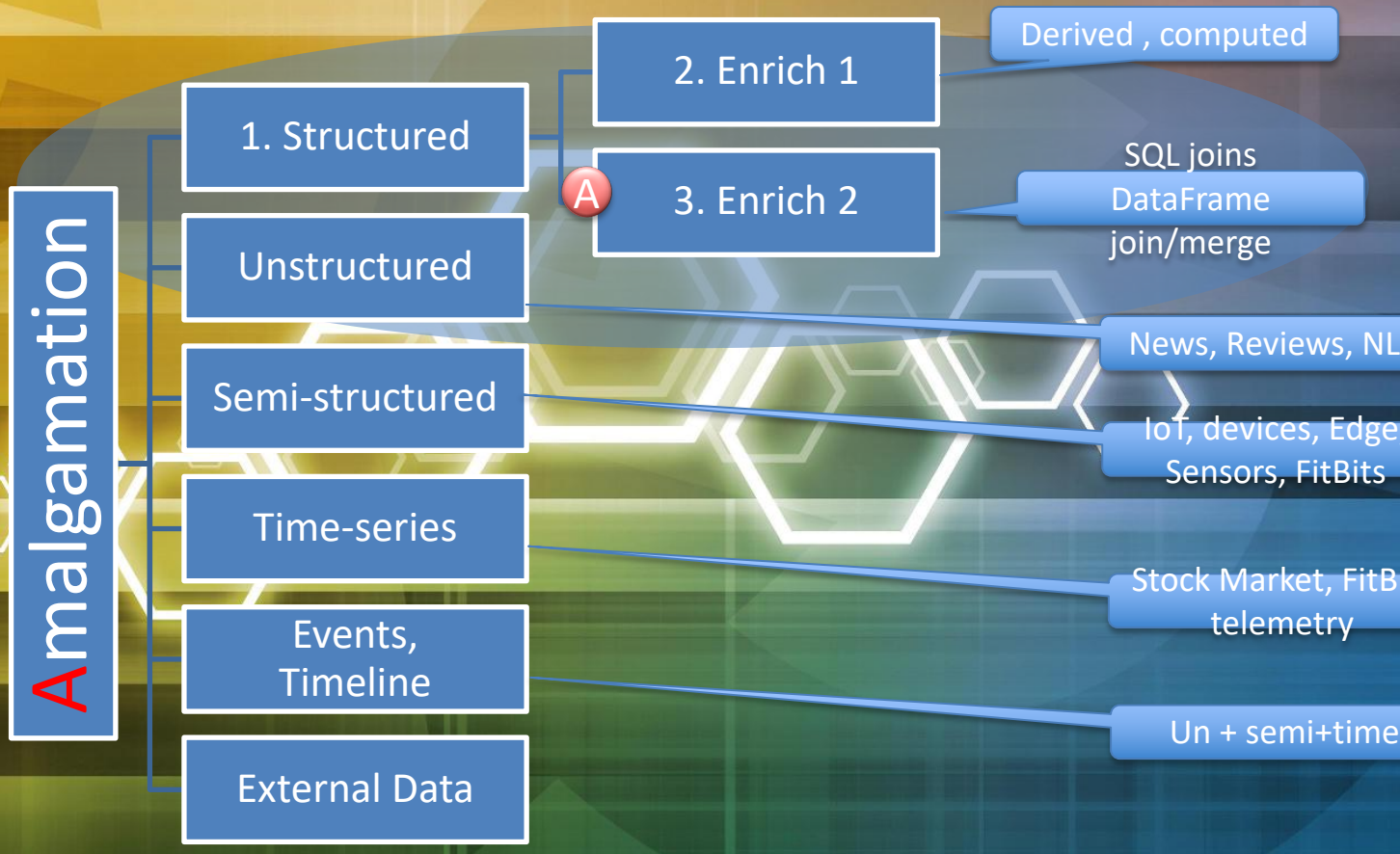
Show Progress



Simple Data Science Process



Amalgamation



Amalgamation

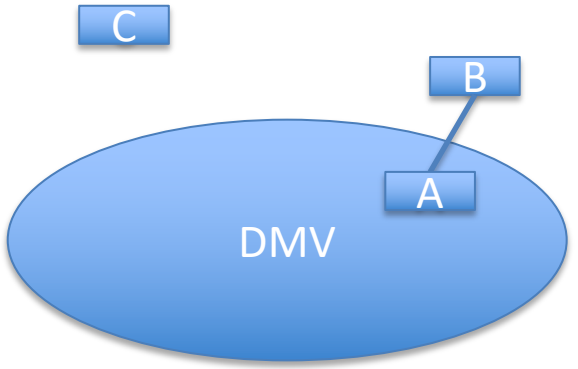
- Can be done in three ways
- 1. linking two datasets by finding a probabilistically "close" value
  - E.g., neighborhood using lat long from one and address from another dataset
- 2. De-biasing a Dataset and rebalancing it
  - BS, comedy, false, true
  - 1000, 500, 300, 200 data points
- 3. Use latent variables in the latent manifold as a way to probabilistically join two tables that would otherwise have *no apparent relationship*.

Embeddings, Distance for Amalgamation

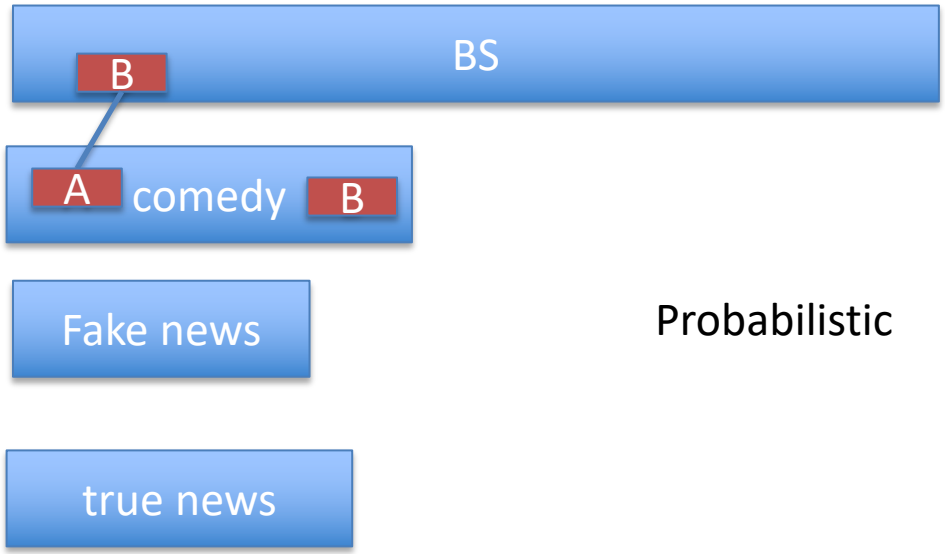
				Neighborhood
Address 123 orchard lane, San jose CA				DMV

Which n(f1) is this address (f2) closest to?

Lat1 -lat2	Long1 – long2	Neighborhood = DMV



De-biasing a Dataset? Using distance (amalgamation)



Probabilistic

Events, External Data, Event Timeline

Events, External Data, Event Timeline

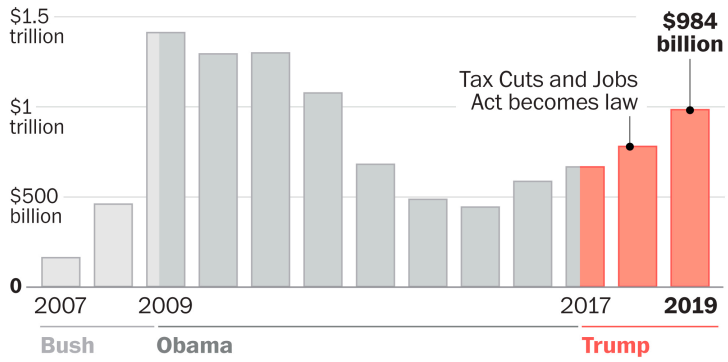
The U.S. **trade deficit** widened to USD 59.8 billion in December of **2018** from an upwardly revised USD 50.3 billion in the previous month and compared with market expectations of a USD 57.9 billion gap. It is the largest **deficit** since October of 2008 as exports declined for the third month and imports recovered. 6 days ago

United States Balance of Trade | 2019 | Data | Chart | Calendar ...

https://tradingeconomics.com/united-states/balance-of-trade

The U.S. budget deficit has more than doubled since 2015

Fiscal-year deficit (The federal fiscal year runs from Oct. 1 to Sept. 30).



Sources: Treasury Department, Office of Management and Budget

THE WASHINGTON POST

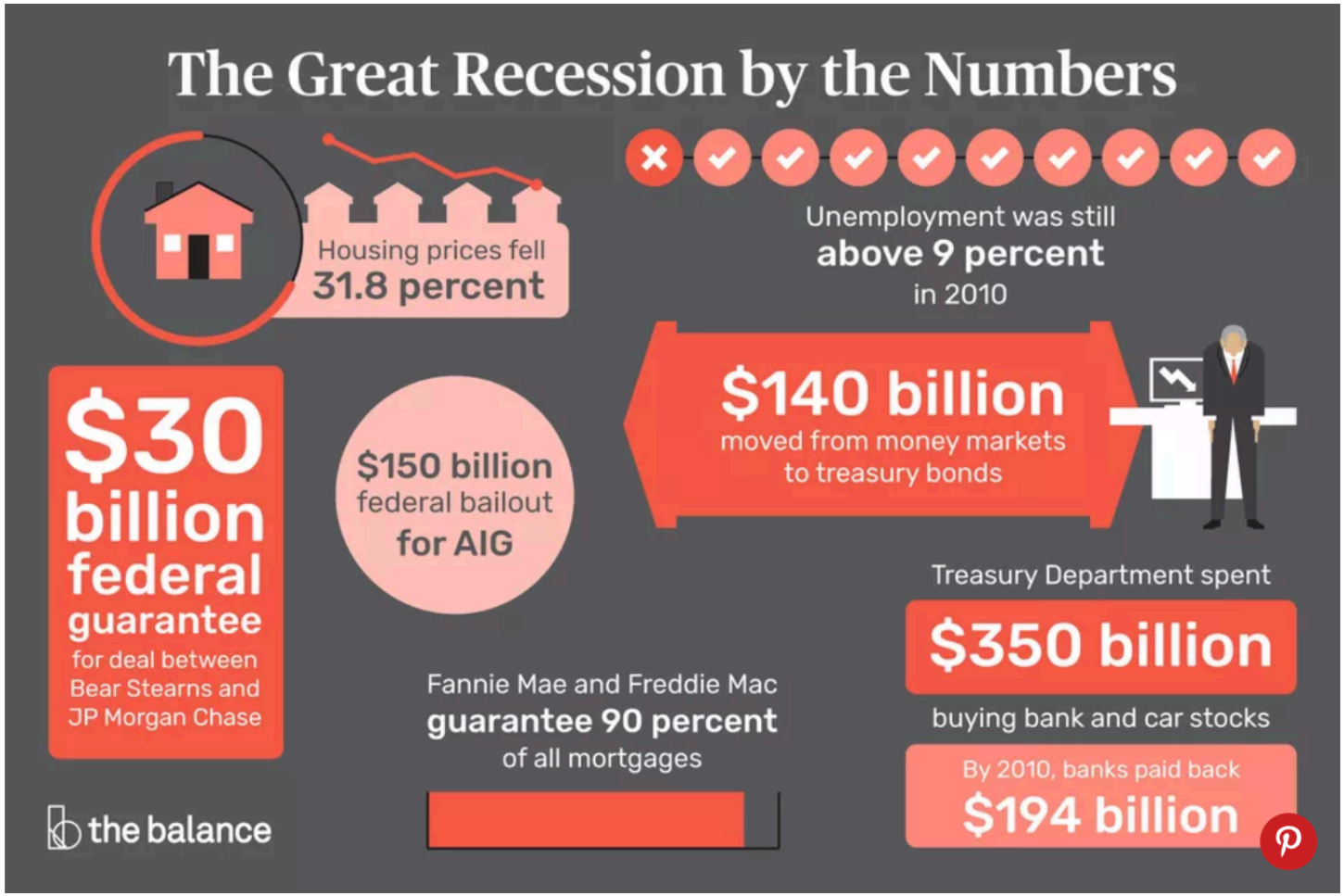
Events, External Data, Event Timeline

The **2008** financial crisis is the worst **economic** disaster since the Great Depression of 1929. It occurred despite Federal Reserve and Treasury Department efforts to prevent it. It led to the Great Recession. That's when housing prices fell 31.8 percent, more than the price plunge during the Depression. Nov 7, 2018

2008 Financial Crisis: Causes, Costs, Could It Reoccur

https://www.thebalance.com/2008-financial-crisis-3305679

Events, External Data, Event Timeline

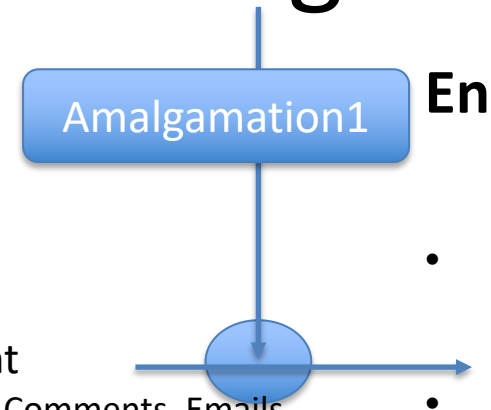


Events, External Data, Event Timeline

- Should I buy an investment property?
  - AirBnB
- Or a restaurant ?

- What does the emerging context (Distillation and Amalgamation and Curation) tell me?

Amalgamation



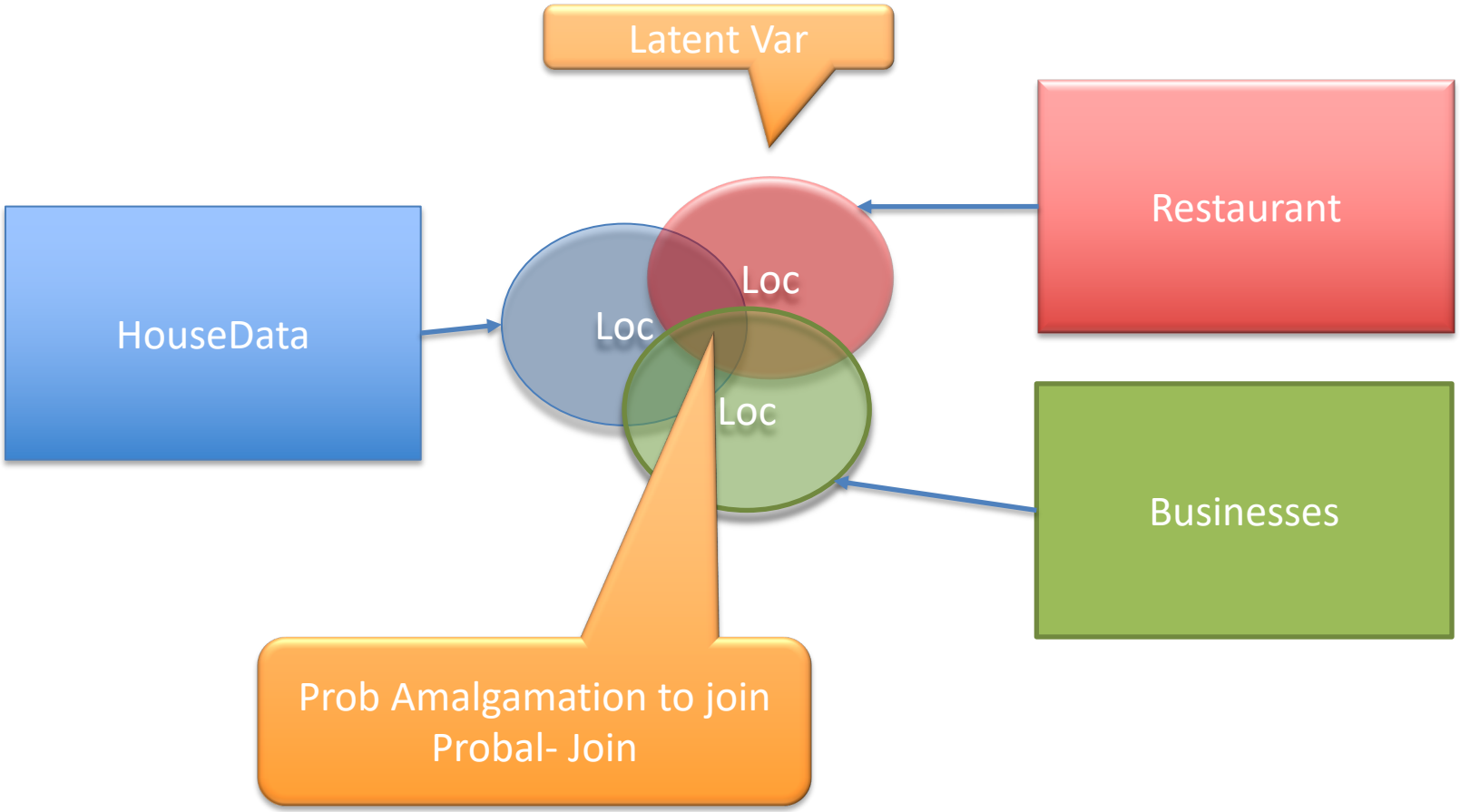
Initial DataSets

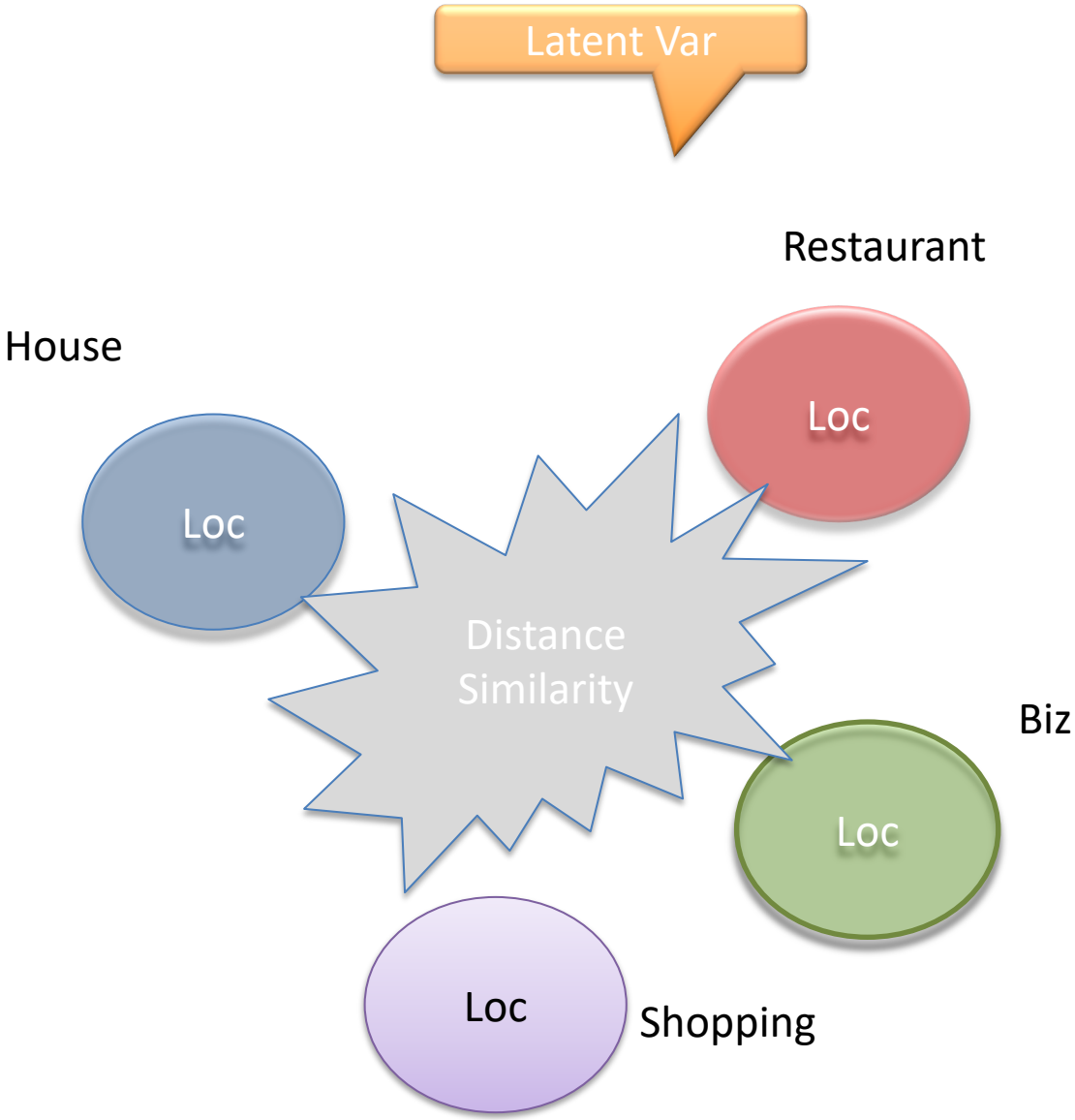
- Structured data
  1. Dataset baseline
- Unstructured Content
  - e.g., Text, Reviews, Comments, Emails, Audio etc.

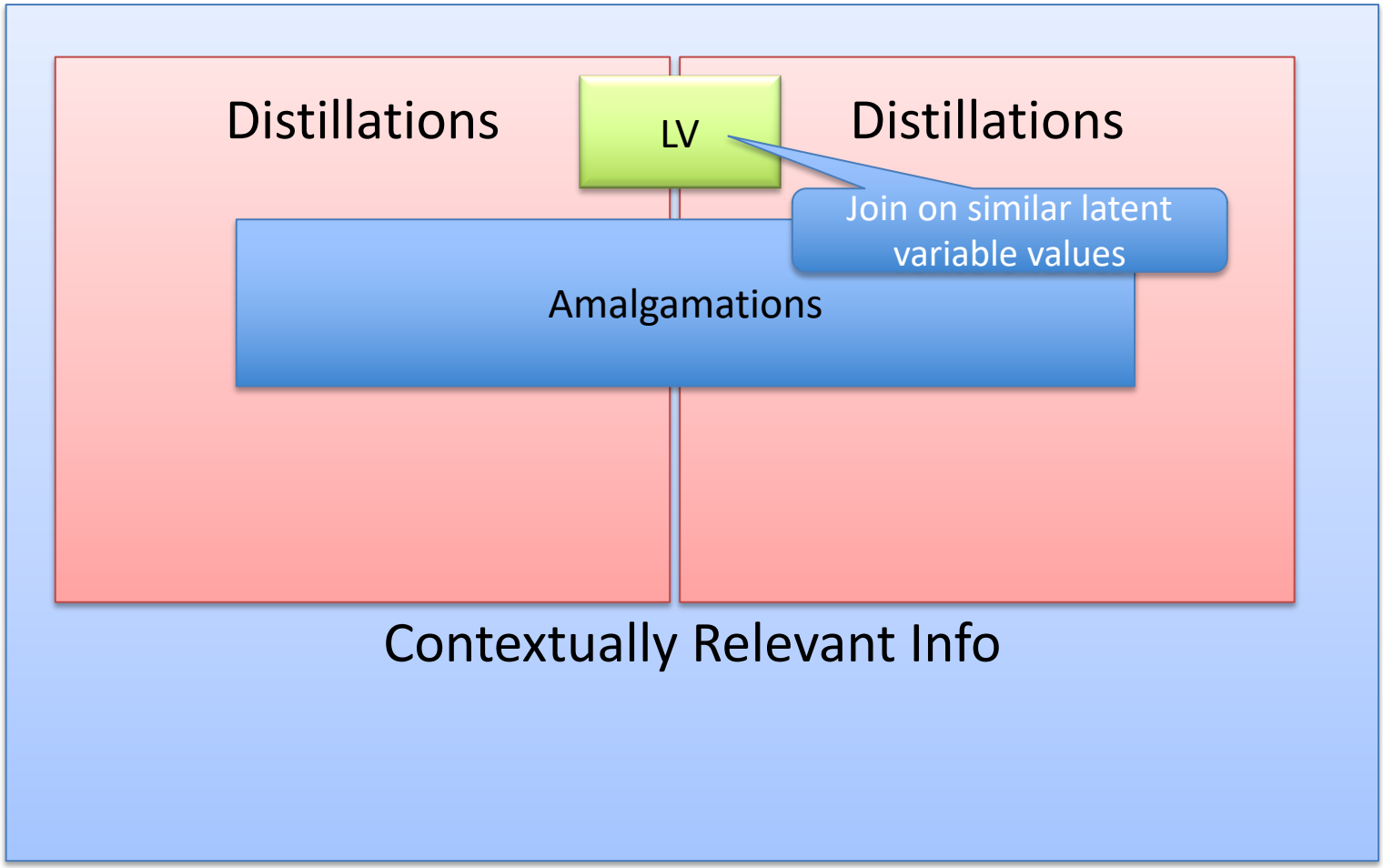
Enrichment DataSets

- Structured Data
  1. Enrichment 1
  2. Enrichment 2
- Times Series
  - Transactions
- Events
  - Everything connected to that event
    - Fires
    - Reviews
    - Emails sent, received
    - Rooms booked
- External Data
  - GeoLocation
  - Counties, neighborhoods
  - Areas
  - Weather

Location Latent Variable




Latent Var
Restaurant
House
Loc
Loc
Distance
Similarity
Biz
Loc
Loc
Shopping



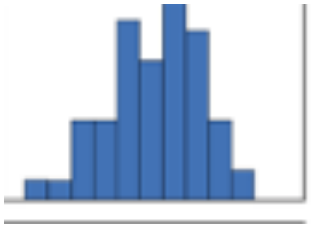
LV

Use latent variables in the latent manifold as a way to probabilistically join two tables that would otherwise have *no apparent relationship*.

Ov1	ov2	Latent
Thai king	thai	[rank index] Proximity to like kind of restaurant + shopping center closeness + residential area closeness + low crime rate
		[price]

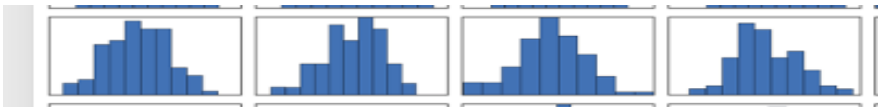
name	cusiine	Prox like	Prox shop	Prox res	poly
		6			
		7			

Polynomial equation :
A1x1 +…+ anxn + a0 = y



Ov1	ov2	Latent
Thai king	thai	[rank index] Proximity to like kind of restaurant + shopping center closeness + residential area closeness + low crime rate
		[price]

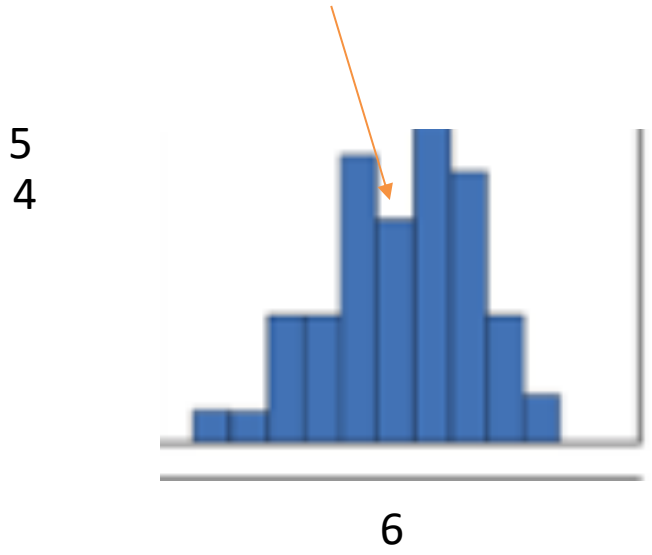
name	cusiine	Prox like	Prox shop	Prox res	poly
???	???	6			
???	???	7			



Polynomial equation :
A1x1 +…+ anxn + a0 = y

Finding the prob distribution of a variable x

X= 6



Shree Gowri Radhakrishna Today at 5:35 PM

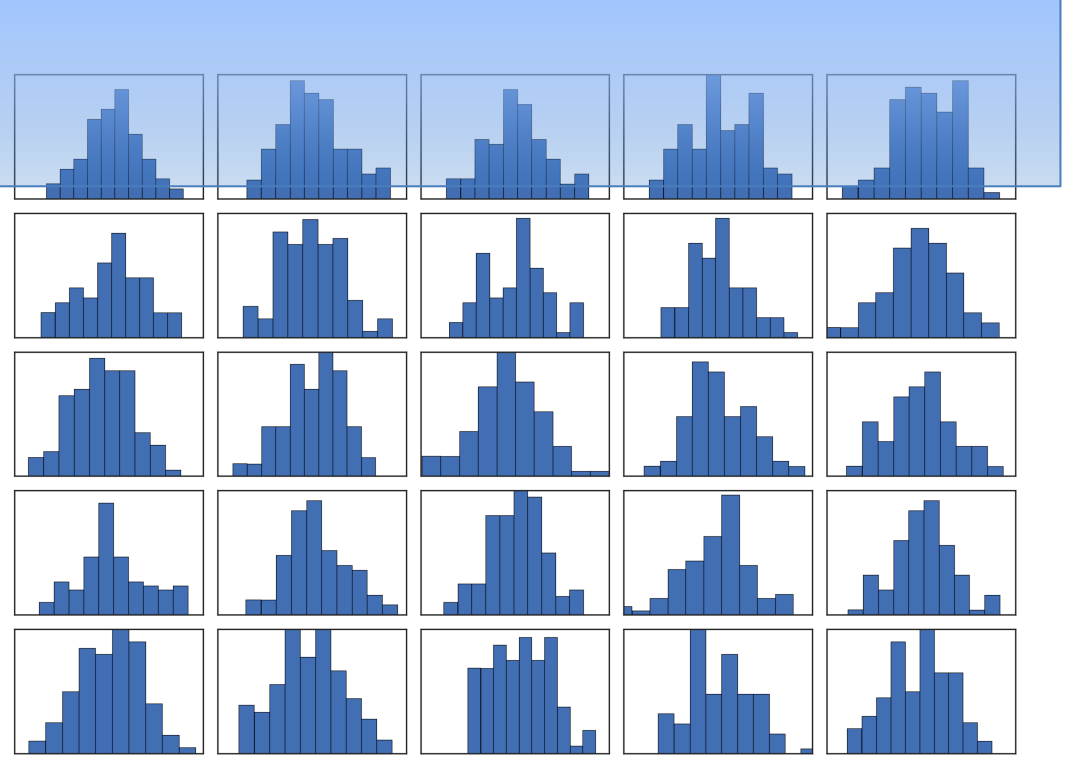
Hi professor, I have a question. How do you amalgamate different kinds of data? Like if we had to merge structured and unstructured or something else, how is the merge happening?

👍 1

1 reply

Dr.Arsanjani < 1 minute ago

Bucket and Map Prob distributions

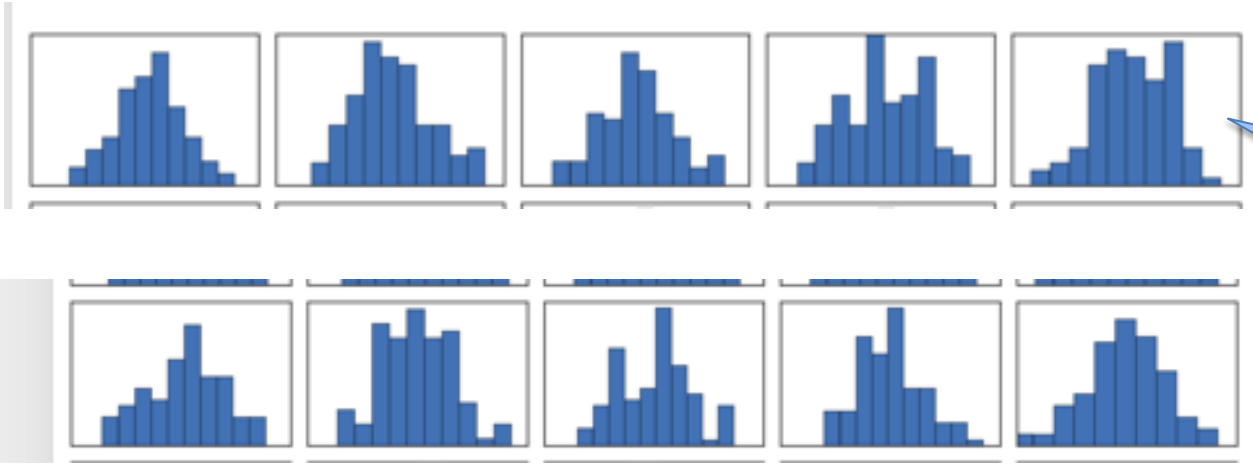


http://work.thaslwanter.at/Stats/html/statsDistributions.html

9000 pop sample
(patients)

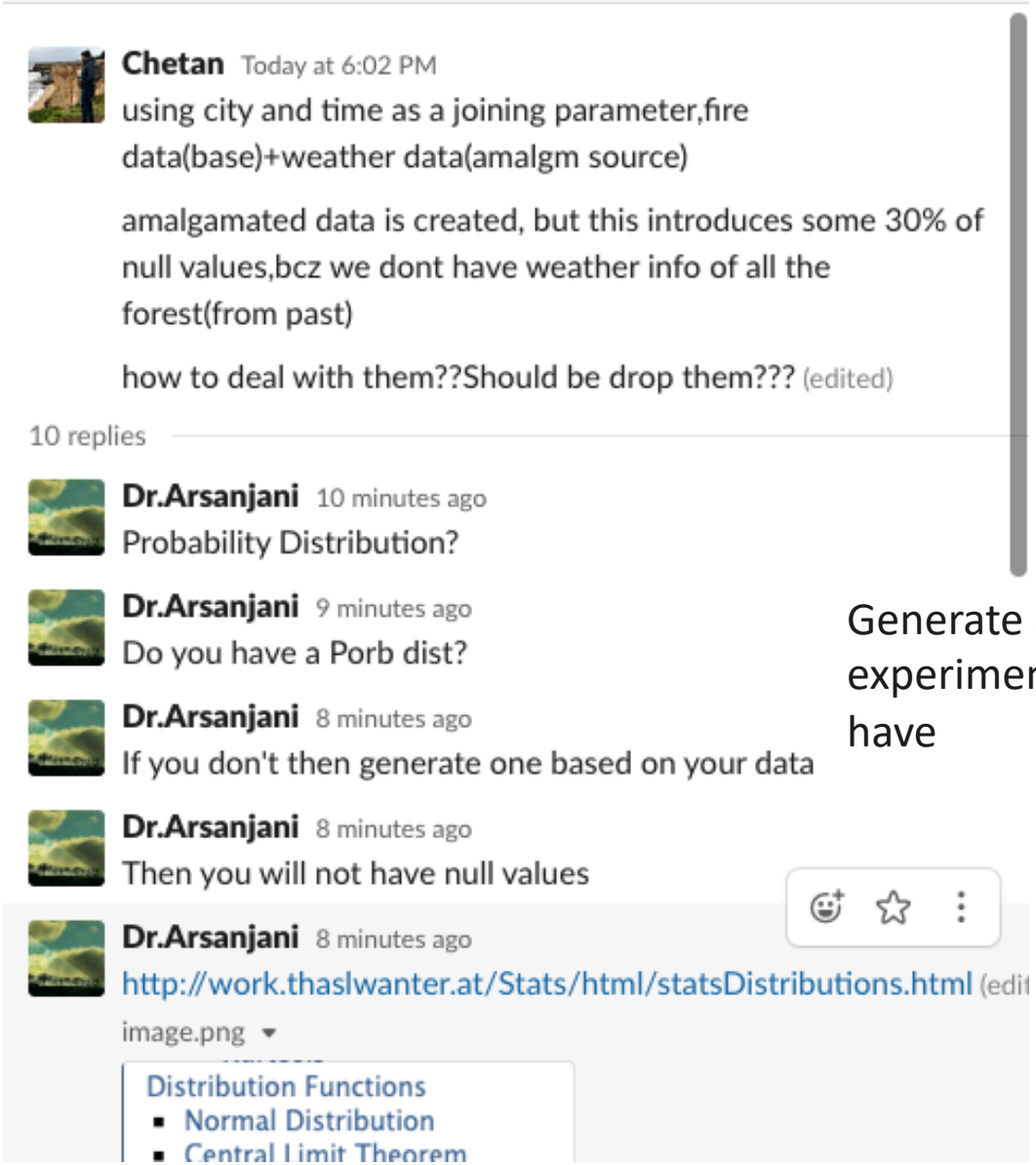
Dataset you find on sample pop maps to 5 types of exercise behavior, activity levels,

Dataset you on fitbit that maps to 5 types of exercise behavior



Prob Dist between the bins

Map the bins that seem to have closer or most similar prob distr & use for amal



Generate a prob distr from the indiv experiments or data points that you currently have

Amalgamation

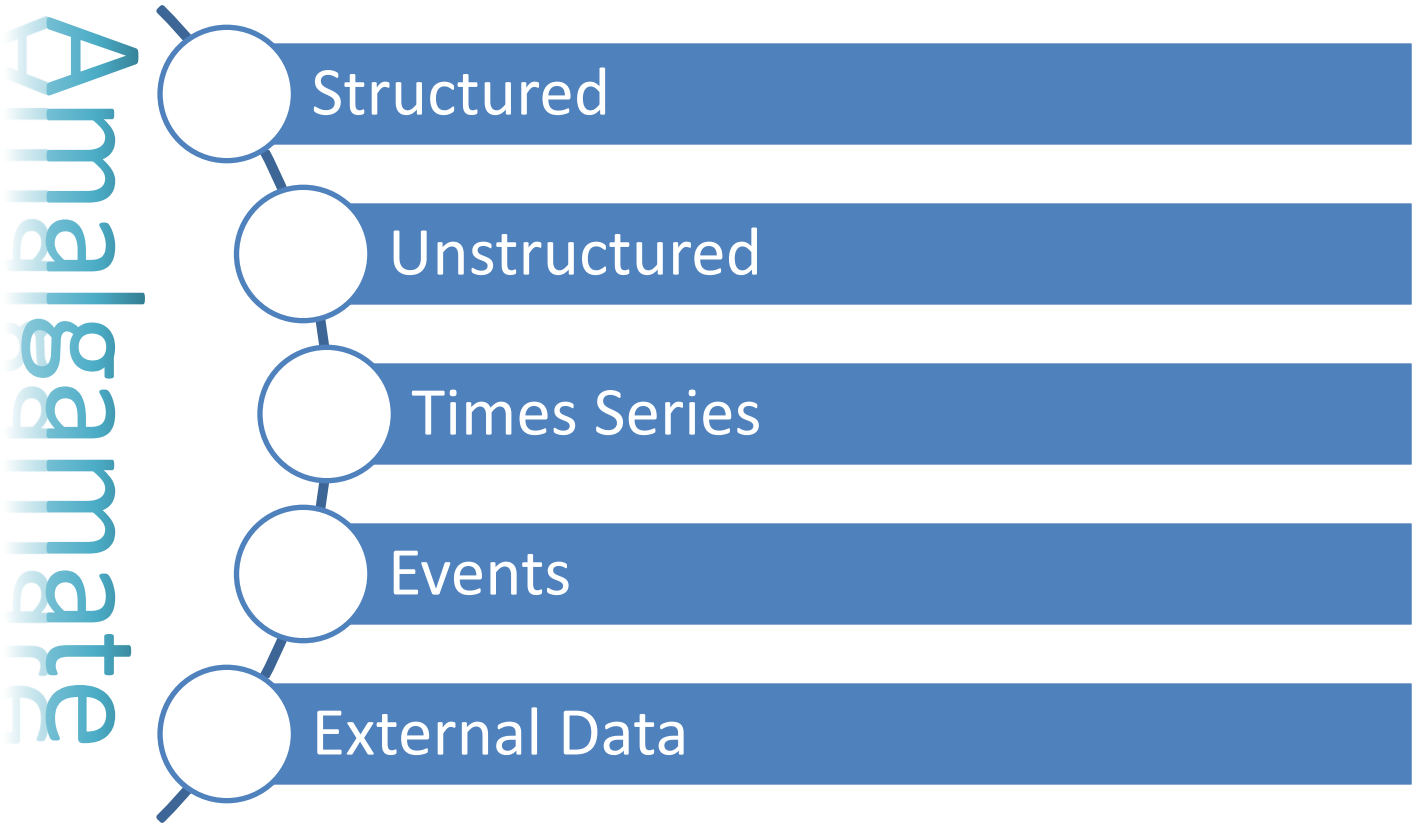
Initial DataSets

- Structured data
  - Looker
  - Gainsight
  - Salesforce
  - Other DBs
- Unstructured Content
  - Voice
  - Video
  - Text

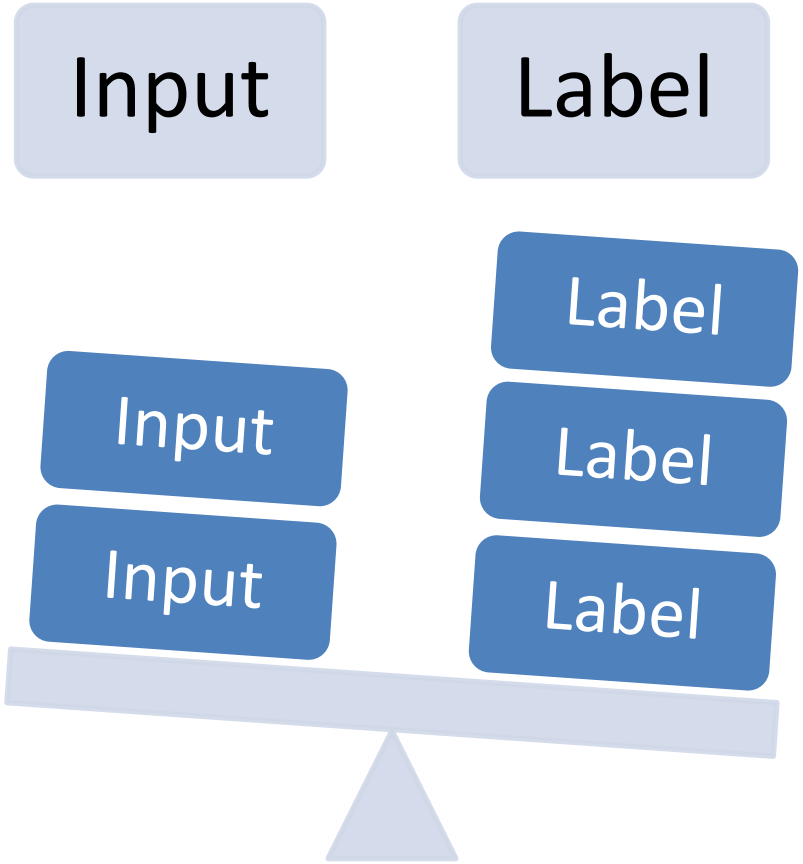
Enrichment DataSets

- Times Series
  - Transactions over time
- Events
  - Everything connected to that event
- External Data
  - G2Crowd
  - Siftery
  - Crunchbase
  - Weather.com

Amalgamation



ICurate.IO



Icurate.io

Cluster

Curate :

- Normalize
- Regularize
- 5-fold cross-validation
- Tag/Annotate

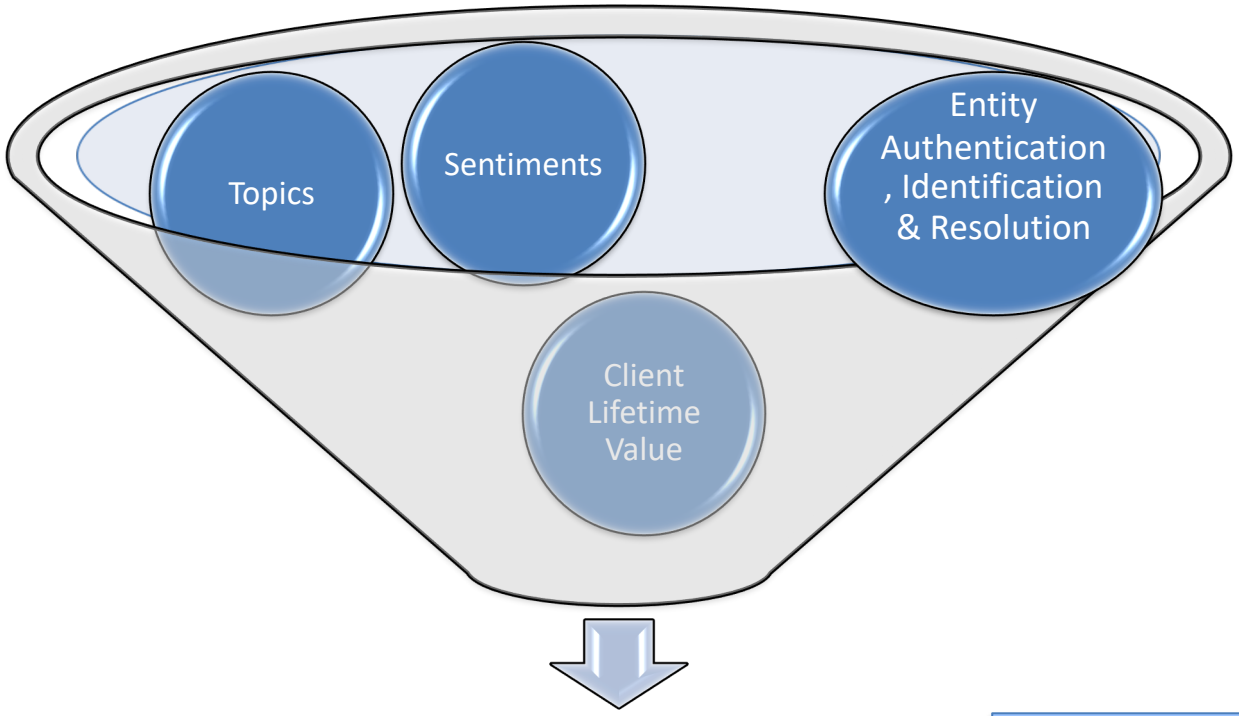
Train regression

Test and Validate Regression model

Classify

- Multivariate

Distillations



Distilled Contextual Embedding

See Distillations Deck

Loss functions

Linear regression and MSE:

$$L(w) = \frac{1}{\ell} \|Xw - y\|^2$$

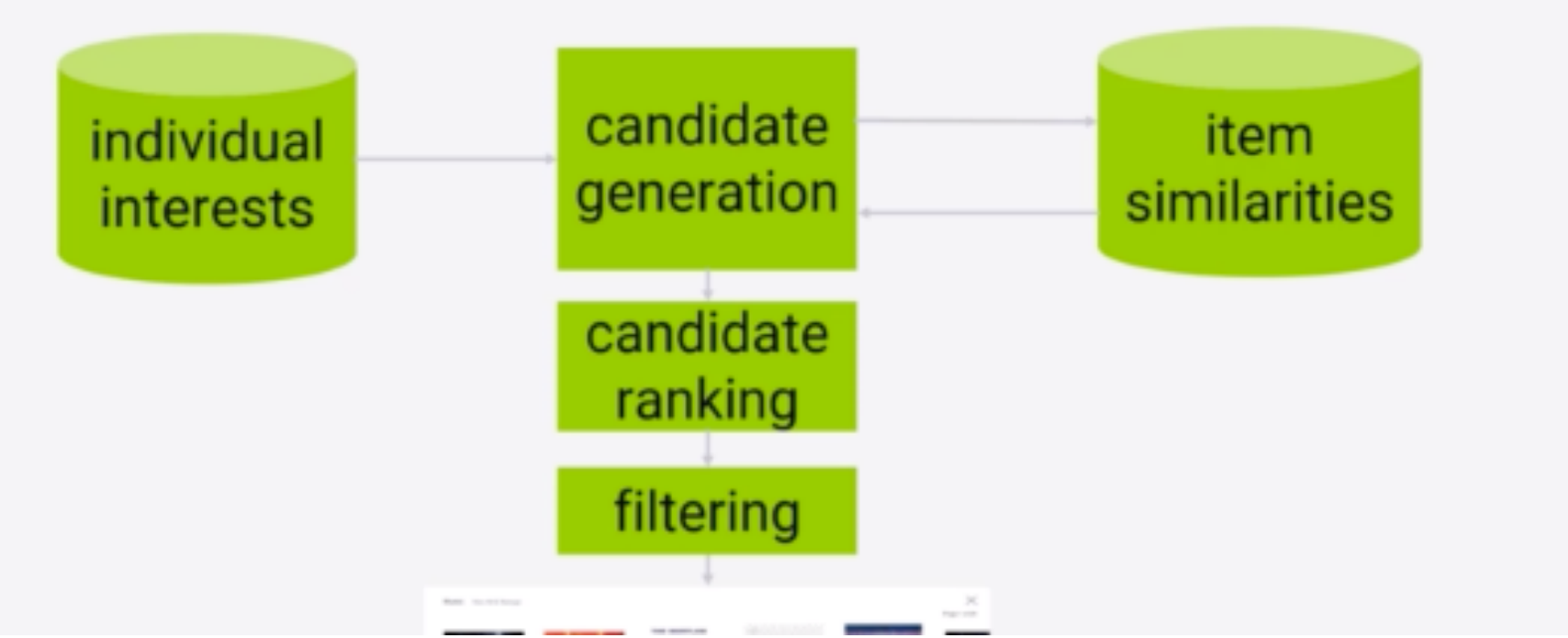
Linear classification and cross-entropy:

$$L(w) = -\sum_{i=1}^{\ell} \sum_{k=1}^{K} [y_i = k] \log \frac{e^{w_k^T x_i}}{\sum_{j=1}^{K} e^{w_j^T x_i}}$$

Cosine Similarity

$$CosSim(x, y) = \frac{\sum_i x_i y_i}{\sqrt{\sum_i x_i^2}\sqrt{\sum_i y_i^2}}$$

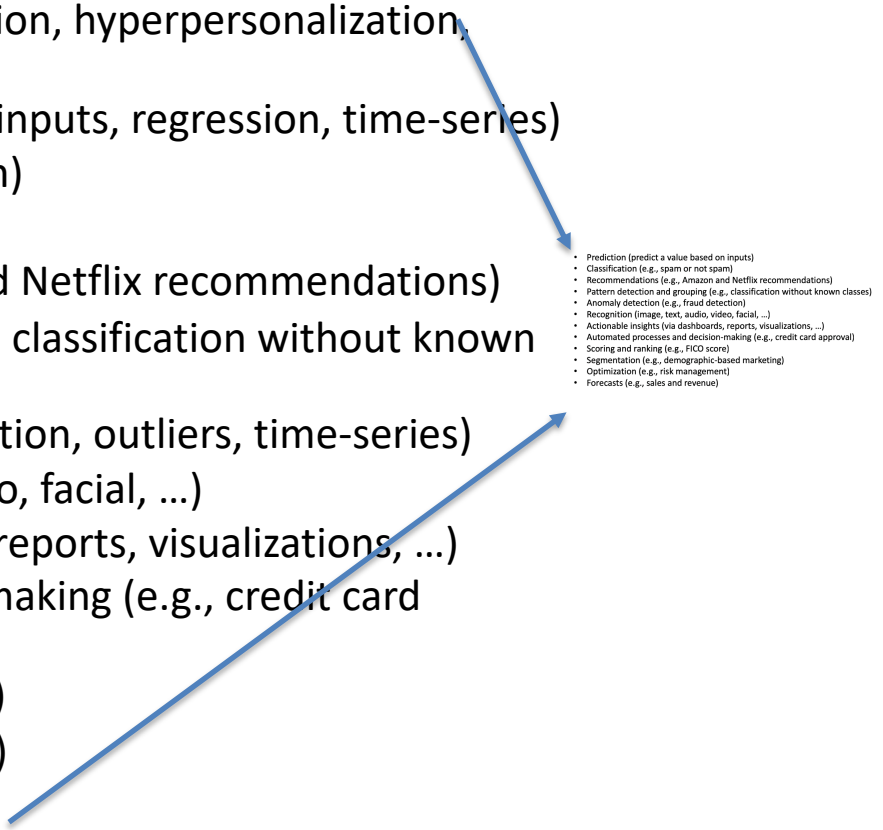
Collaborative Filtering



Things we do in ML

- Segmentation (e.g., microsegmentation, hyperpersonalization, demographic-based marketing)
- Prediction (predict a value based on inputs, regression, time-series)
- Classification (e.g., spam or not spam)

- Recommendations (e.g., Amazon and Netflix recommendations)
- Pattern detection and grouping (e.g., classification without known classes,CNNs)
- Anomaly detection (e.g., fraud detection, outliers, time-series)
- Recognition (image, text, audio, video, facial, …)
- Actionable insights (via dashboards, reports, visualizations, …)
- Automated processes and decision-making (e.g., credit card approval)
- Scoring and ranking (e.g., FICO score)
- Optimization (e.g., risk management)
- Forecasts (e.g., sales and revenue)



Events

- **News**

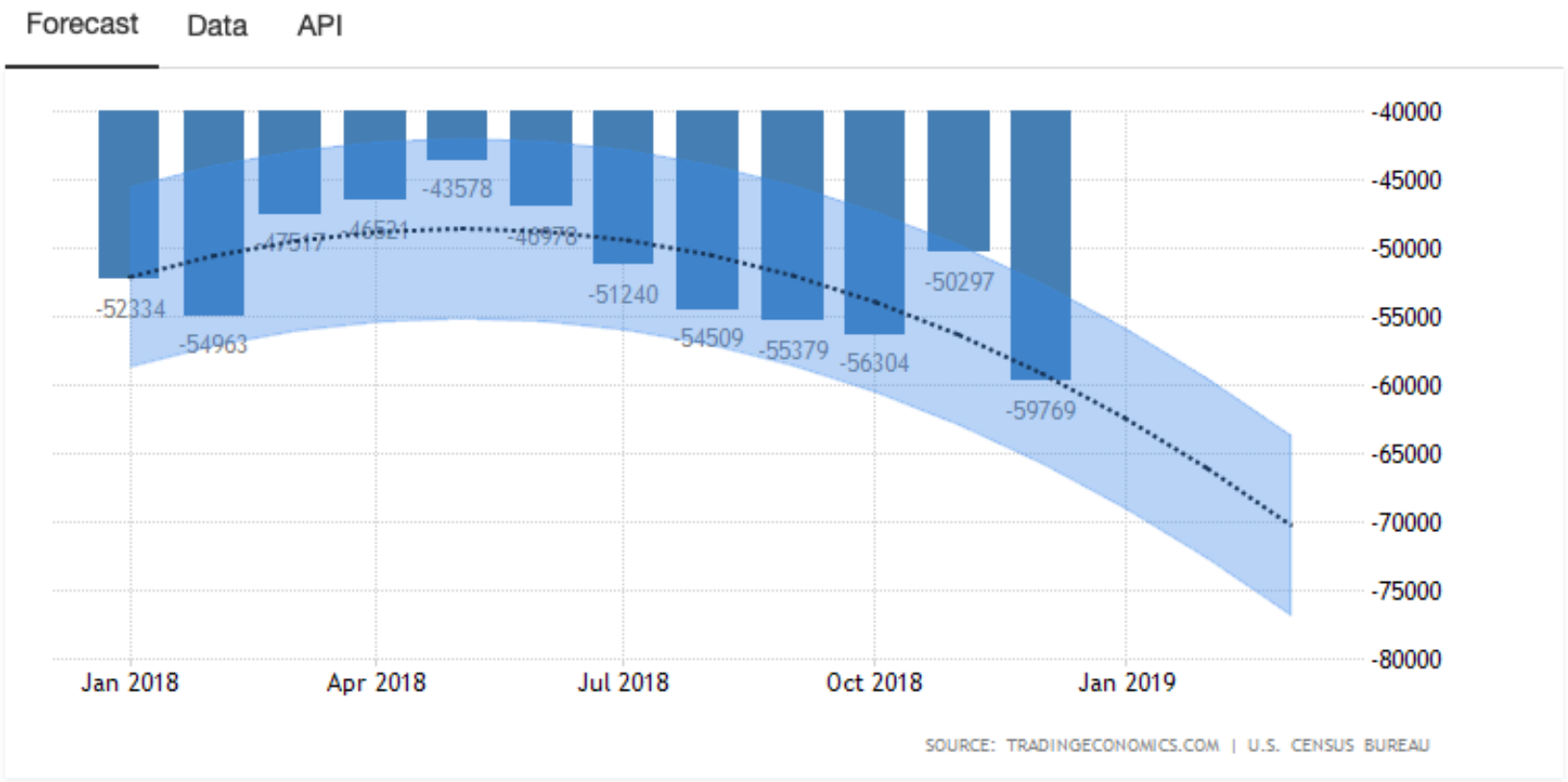
  US Trade Gap Lowest in 5 Months
  The U.S. trade deficit narrowed to USD 49.3 billion in November of 2018 from an upwardly revised USD 55.7 billion in the previous month and compared with market expectations of a USD 54 billion gap. It is the lowest deficit in five months as imports plunged the most since March of 2016 from a record high value reached in the previous month.
  Published on 2019-02-06

  US Trade Deficit Reaches 10-Year High
  The U.S. trade deficit widened to USD 55.5 billion in October of 2018 from an upwardly revised USD 54.6 billion in the previous month and compared with market expectations of a USD 54.9 billion gap. It is the highest deficit since October of 2008 as lower soybean sales weighed down on exports and imports reached a new record high.
  Published on 2018-12-06

  US Trade Deficit Highest in 7 Months
  The U.S. trade deficit widened to USD 54.0 billion in September 2018 from an upwardly revised USD 53.3 billion in the previous month and compared with market expectations of a USD 53.6 billion gap. It is the highest deficit in 7 months as imports rose to a record high.
  Published on 2018-11-02

  US Trade Deficit Jumps to 6 Month High
  The U.S. trade deficit increased to a six-month high of USD35.2 billion in August as exports dropped further amid declining soybean shipments and imports hit a record high amid stronger demand for cars, industrial supplies and petroleum.
  Published on 2018-10-05

Regression


Forecast
Data
API
-40000
-45000
-50000
-55000
-60000
-65000
-70000
-75000
-80000
-52334
-54963
-47517
-46521
-43578
-46978
-51240
-54509
-55379
-56304
-50297
-59769
Jan 2018
Apr 2018
Jul 2018
Oct 2018
Jan 2019
SOURCE: TRADINGECONOMICS.COM | U.S. CENSUS BUREAU

Data Pipeline

Calendar	GMT		Actual	Previous	Consensus	TEForecast
2018-11-02	12:30 PM	Balance of Trade	$-54B	$-53.3B	$-53.6B	$ -53B
2018-12-06	01:30 PM	Balance of Trade	$-55.5B	$-54.6B	$-54.9B	$ -55.4B
2019-02-06	01:30 PM	Balance of Trade	$-49.3B	$-55.7B	$-54B	$-53B
2019-03-06	01:30 PM	Balance of Trade	$-59.8B	$-50.3B	$-57.9B	$-56B
2019-03-27	12:30 PM	Balance of Trade		$-59.8B		
2019-04-17	12:30 PM	Balance of Trade				
2019-05-09	12:30 PM	Balance of Trade				

+

Distillations: Week 7

Dr. Ali Arsanjani

Machine Learning

10/8/20

Use Big Data driven PREDICTIVE ANALYTICS to Optimize Sales

- Context helps refine and focus advanced predictive analytics solutions to
- prioritize and optimize the greatest potential for
  - high-volume sales,
  - reduce churn and
  - increase relevant upsell

- Organizations can optimize
- Marketing spend and account management to
- increase estimated revenue

How GM Uses BIG DATA to Generate Sales

- General Motors combined big data, analytics, and GIS to model dealership performance.
- It enabled dealers from around the nation to view local demographics, location characteristics, and regional differences to providing a dealership the ability to compare their performance to actual results.
- **Driving for a deal**
- Instead of approaching the marketing arena, which GM budgets around $2 billion each year, they *decided to conduct an analyses to determine the types of households that will buy the various automobiles within its portfolio*.
- By feeding detailed demographic and spatial data to marketing, they were able ***direct its ad spend*** to the right departments.
- Leverage DeepContext , “ask me how I remember tomorrow?” ™

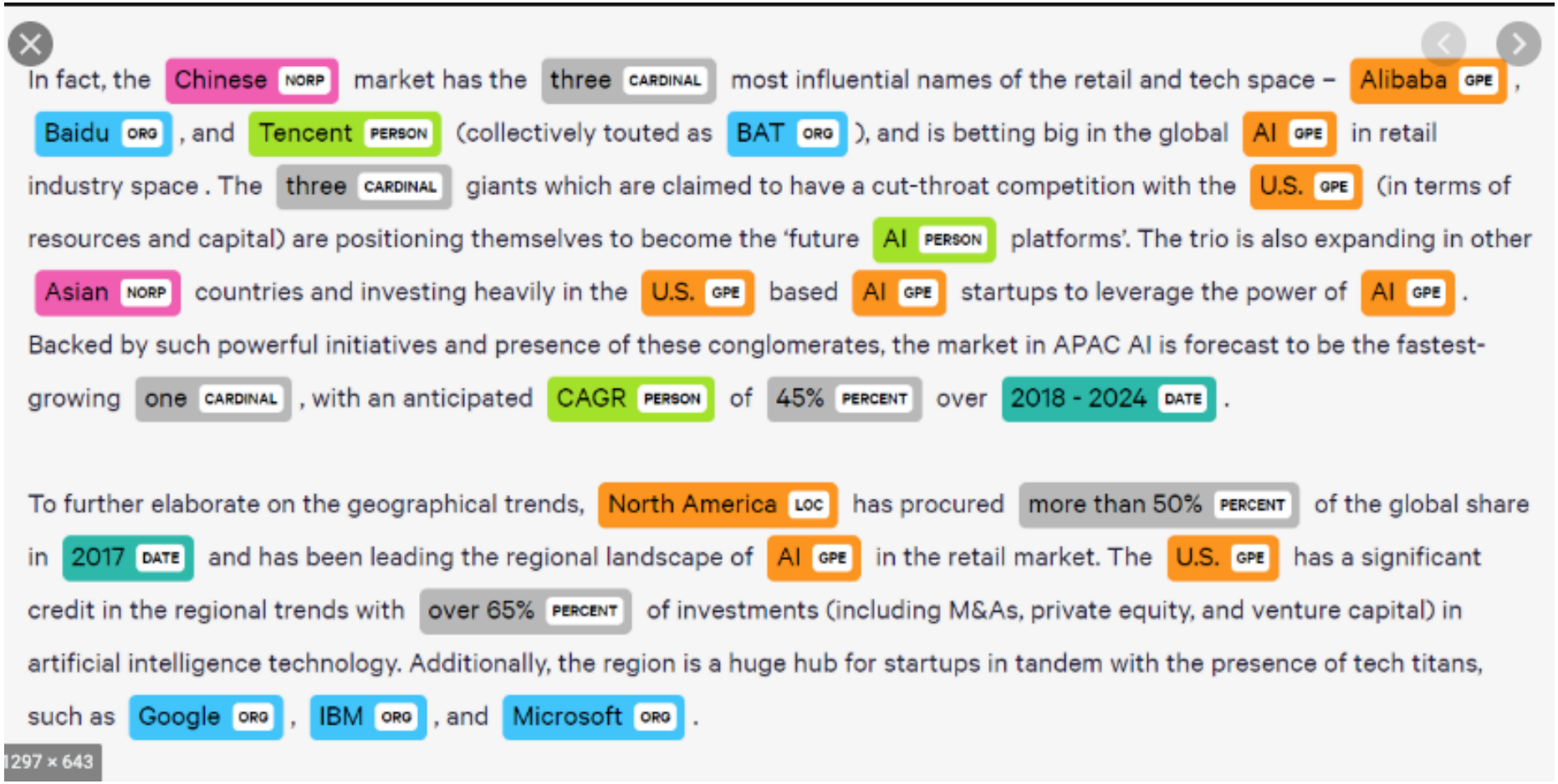
Deep Context Data Science Life-cycle

- Data Selection.
  - Find one baseline dataset,
- Data Narrative.
  - Figure out your data narrative
- Algorithm Selection
  - Which algorithms (clustering, regression, classification) are a better fit for your data, and to fulfill your narrative?
- Latent Variables, Models, Manifolds
  - Identify at least one Latent Manifold, which variables should you use?
  - Enrich data in the next step to feed into your latent manifold, to help with explaining results and contributing to your Data Narrative
- Data Enrichment .
  - Find 2 other datasets that will help deepen insights and refine results of regression accuracy
- Amalgamation
  - Apply amalgamation techniques, identify and report back on your amalgamation and how it possibly increased your accuracy, R2, F1, recall, precision, RSME, etc.
- Data Distillation
  - Can you distill your data?
- Curate your data for deeper accuracy, insight
  - Supervised learning, reinforcement learning

Distillations

1. Customer Identity
   - (basic, lookup: e.g., caller id → pull up record in CRM)
2. Entity Resolution
   1. Amalgamation:
      1. Embeddings, Cosine Distance (words that have meanings close to one another, or occur in similar **contexts**)
      2. Literals, Euclidean distance (e.g., geolocation to find common communities, areas, etc)
   2. E.g., Senzing.com
3. Customer Lifetime Value (complex), Customer Rank (simple)
   - E.g, simple : how much have they purchased to date? How much do we anticipate (regression) they will purchase in the future based on prior purchases (time-series)? What is their propensity to buy? To convert to a customer ? What is their propensity for an upsell?
4. Sentiment Analysis
   1. Sentiment a la *Vader*
   2. Tone Analysis
   3. Personality Insights
5. Topics
   - LDA, LSA, variations
     1. Attention based LSTM with n-grams
6. Requests/Intent
   - Use topics to unearth actual requests customers are making, what are they asking when they call, text, email?
7. Time lines
   - Construct an event timeline of the customer behavior, patient case, insurance claim, stock symbol relative to market, realestate prices, etc.
8. Locations
   - Geolocation, community, lat-long, zipcode, etc.
9. Entities and Relationships Extraction (Named Entity Recognition (NER) Extraction)
   - For knowledge graph construction
10. Dictionary,
    - construct a lexicon, taxonomy, of jargon

NER: Tags Known Entities with Meta-data

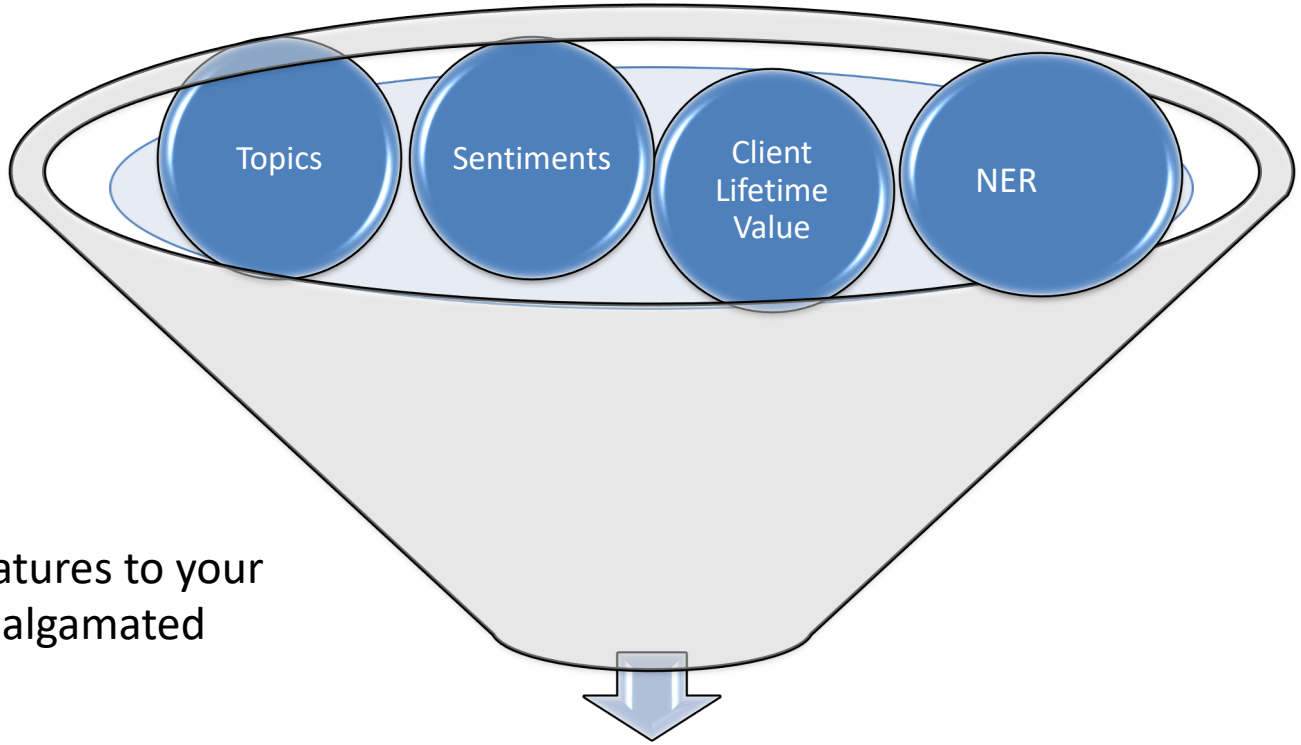


Adding new entities and meta data

- That is not known by the NER
- Custom NER
  - Dict : { “<entity word>” : “<label>”}
- Topics are input for your dict

Distillations

NLP



Add new features to your existing, amalgamated dataset

Distilled Contextual Embedding

Add more accuracy, precision, recall, f1 , RMSE, CM


Machine Learning Week 7